Simulating Early Word Learning in Situated Connectionist Agents

Felix Hill (felixhill@google.com)

Stephen Clark (clarkstephen@google.com)

Karl Moritz Hermann

Phil Blunsom

DeepMind, London, UK

Abstract

Recent advances in Deep Learning (DL) and Reinforcement Learning (RL) make it possible to train neural network agents with raw, first-person visual perception to execute language-like instructions in 3D simulated worlds. Here, we investigate the application of such deep RL agents as cognitive models, specifically as models of infant word learning. We first develop a simple neural network-based language learning agent, trained via policy-gradient methods, which can interpret single-word instructions in a simulated 3D world. Taking inspiration from experimental paradigms in developmental psychology, we run various controlled simulations with the artificial agent, exploring the conditions in which established human biases and learning effects emerge, and propose a novel method for visualising and interpreting semantic representations in the agent. The results highlight the potential utility, and some limitations, of applying state-of-the-art learning agents and simulated environments to model human cognition.

Keywords: early word learning; neural networks; situated artificial agents; 3D environments; word learning biases

Introduction

The learning challenge faced by children acquiring their first words has long fascinated philosophers, linguists and cognitive scientists (Bloom, 2000). To start making sense of language, an infant must induce structure in a stream of continuous visual input, reconcile this structure with consistencies in the linguistic observations, store this knowledge in memory, and apply it to inform decisions about how best to respond to new utterances.

Many neural network models also overcome a learning task that is – to varying degrees – analogous to early human word learning. Image classification tasks such as ImageNet require models to induce discrete semantic classes, aligned to words, from unstructured pixel representations of large quantities of photographs (Krizhevsky, Sutskever, & Hinton, 2012). Visual question answering systems (e.g. Antol et al. (2015)) must reconcile raw images with sequences of symbols, in the form of natural language questions, in order to predict lexical or phrasal answers. More recently, situated artificial agents have been developed that learn to understand sequences of words not only in terms of the contemporaneous raw visual input, but also in terms of past visual input and the actions required to execute an appropriate motor response (e.g. Oh, Singh, Lee, and Kohli (2017), Hill et al. (2020)). The most advanced such agents learn to execute a range of phrasal instructions, such as *find the green object in the red room*, in a continous, simulated 3D world. To solve these tasks, an agent must execute long sequences of (comparatively) fine-grained actions, conditioned on the available language string and active first-person visual perception.

Here, we consider the utility of deep RL agents, trained and tested in a 3D game world, as models of human cognition; specifically of early word learning. The customisable nature of the world, including a limited set of objects, properties, and symbolic linguistic stimuli (Fig. 1B), allows us to replicate several well-known experimental paradigms normally applied with human learners. In a typical experimental episode, the agent is presented with a single word and two objects in a room. It must move by choosing between eight motor actions, viewing the objects from different perspectives until it can determine which one best reflects the meaning of the word.[1] It receives a scalar positive reward if it selects the correct object by moving towards and bumping into it.

We show that, under certain training conditions, our agent comes to exhibit various aspects of early word learning. First, the agent successfully learns a vocabulary of words from different semantic classes, and we study the dynamics of this process. We show that the rate at which the agent acquires new words increases rapidly after an initial slow period, an effect matching the human vocabulary spurt (Plunkett, Sinha, Møller, & Strandsby, 1992). We also propose two ways to speed up word learning: moderating the agent's experience according to a curriculum (Elman, 1993) and an auxiliary learning objective reinforcing the association between words and the agent's replayed visual experience. Second, we investigate whether the agent exhibits a shape or colour bias (MacWhinney, 1999; Regier, 2003). And finally, for a better view of how the model processes information at the algorithmic level, we develop a novel method for dynamically visualising how different word types stimulate activations in different parts of its architecture. Taken together, these simulations illustrate how the combination of DL and RL together may be a fruitful, if imperfect, basis for building holistic simulations of human semantic cognition.

[1] There is clearly more to knowing the meaning of a word than being able to identify an appropriate referent, but we are inspired by how infants initially learn to identify objects.

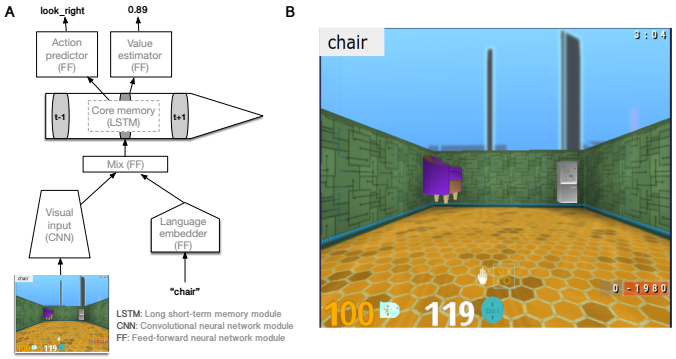

Figure 1: **A:** Schematic agent architecture. **B:** An example of the word learning environment.

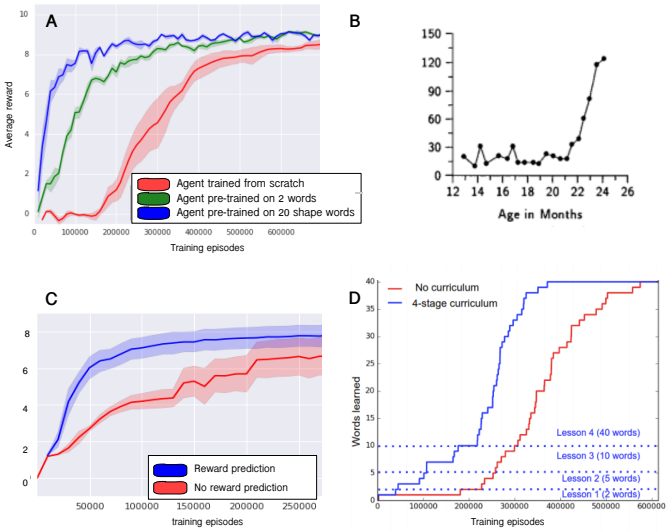

Figure 2: **A** Word learning trajectories for the agent. **B** The acceleration of vocabulary size in an infant. **C** The effect of reward-prediction auxiliary loss on learning speed. **D** Word learning trajectories for an agent following a curriculum.

A 3D world for language learning

We fix the overall layout of the world (a rectangular room), the range of positions in which the agent begins an episode (near the back of the room), the locations that objects can occupy (two positions at the front), a list of objects that can appear, the relative frequency of each object appearing, and rewards for selecting a certain object given a particular instruction word. The environment engine is then responsible for randomly instantiating episodes that satisfy these constraints together with corresponding instruction words. Even with this relatively constrained level specification, there are a huge number of unique episodes that the agent can encounter during training, each involving different object shapes, colours, patterns, shades, and relative positions.

A situated word-learning agent

Our agent (Figure 1A), combines standard modules for processing symbolic input (an embedding layer) and visual input (a three-layer convolutional network). At each time step t, the visual input v_t (a $3 \times 84 \times 84$ tensor of floating point RGB pixel values) is encoded by the convolutional *vision module* into a 3136 (= 64 feature maps $\times$ 49 locations) dimensional embedding, and a *language module* embeds the instruction word l_t into a 128 dimensional embedding. A *mixing module* determines how these signals are combined before they are passed to an LSTM *core memory*. In this work, the mixing module is simply a feedforward linear layer that maps the concatenation of the output from the vision and language modules to a 256-dimensional embedding. The language module is a simple linear lookup weight matrix (since the instruction consists of one word) applied to one-hot encodings of the input words. Thus, prior to learning, the model has no prior information about the correct reference, or word class, for the different types of words that it experiences.

The 256-dim hidden state s_t of the core memory Long Short Term Memory (LSTM) module is fed to an action predictor (a fully-connected layer plus softmax), which computes the policy, a probability distribution over possible motor actions $\pi(a_t|s_t)$, and a state-value function estimator $Val(s_t)$, which computes a scalar estimate of the agent state-value function (the expected discounted future return). This value estimate is used to compute a baseline for the return in the asynchronous advantage actor-critic (A3C) policy-gradient algorithm (Mnih et al., 2016), which determines weight updates in the network in conjunction with the RMSProp optimiser (Tieleman & Hinton, 2012). The weights in the visual, language and core memory modules are trained end-to-end.

Word learning dynamics

In our first simulation, we randomly initialized all of the weights in the agent network, and then trained it on episodes with instruction words referring to the *shape*, *colour*, *pattern*, *relative shade* or *position* of objects. There were 40 shape words, e.g. "pencil"; 10 colour words, e.g. "blue"; 2 pattern words, e.g. "striped"; 2 shade words, e.g. "darker"; and 2 direction words, e.g. "left". An instruction such as "blue" would mean find the blue object, and the agent would be rewarded by bumping into the blue object at the other end of the room. The instruction word in each episode unambiguously specified one of the two target objects, but other unimportant aspects of the environment could vary maximally. Thus, shape-word instructions could refer to objects of any colour, colour-words to objects of any shape, and so on. The agent received a reward of $+10$ if it bumped into the correct object, -10 if it bumped into the wrong object, and 0 if the maximum number of timesteps was reached. All words appeared with equal frequency during training.

We found that the agent slowly learned to respond correctly to the words it was presented with, but at some point the rate of word learning accelerated rapidly (Fig. 2A, red curve). This effect is observed in both young infant learners (Nazzi & Bertoncini, 2003) and (supervised) connectionist simula-

tions of word learning (Fig. 2B, as recorded by Plunkett et al. (1992)). Our results show that the effect persists when such networks are trained with RL algorithms from raw pixel input. By the end of successful training, the agent was able to walk directly up to the two objects and reliably identify the appropriate referent.[2]

For our agents, some of the delay in the onset of word learning can be explained by the need to acquire relatively language-agnostic capacities such as useful sequences of motor actions or the distinction between objects and walls. However, some of the acceleration seems also to derive from the accruing semantic knowledge. To demonstrate this, we compared word learning speeds in an agent with prior knowledge of 2 words to an agent with knowledge of 20 words (Fig. 2A, green and blue curves). The prior knowledge was provided by training the agent on the word-learning task, as described above, but restricting the vocabulary to 2 and 20 words. So in both cases the agent has learned to "see" and move, but the agent pre-trained on 20 words learned new words more quickly. This effect accords with accounts of human development that emphasise how learning becomes easier the more the language learner knows (Bates & MacWhinney, 1987).

We also explored ways to reduce the number of rewarded training episodes before word learning onset, in the form of a curriculum. We found one way to achieve this by moderating the scope of the learning challenge faced by the agent initially, before later expanding its experience once word learning had started. Specifically, we trained the agent to learn the meaning of the 40 shape words under two conditions. In one condition, the agent was presented with the 40 words (together with corresponding target and confounding objects) with uniform random frequency throughout training. In another condition, the agent was only presented with a subset of the 40 words (with uniform random frequency across that subset) until these were mastered (as indicated by an average reward of $9.8/10$ over 1000 consecutive trials), at which point this subset was expanded to include more words. So the stimuli are initially constrained to a two-word subset $S_1, S_1 \subset S$, until the agent learns both words, then extended to a 5-word subset $S_2, S_1 \subset S_2 \subset S$, then a 10-word subset $S_3, S_2 \subset S_3 \subset S$, until finally being exposed to all 40 words in S. As shown in Fig. 2D, the agent following the curriculum reached 40 words faster than the agent confronted immediately with a large set of new words. This effect accords with the idea that early exposure to simple linguistic input helps child language acquisition (Fernald, Thorpe, & Marchman, 2010), and with *curriculum effects* observed when training neural networks on text-based language data (Elman, 1993; Bengio, Louradour, Collobert, & Weston, 2009).

We found a further way to reduce the number of episodes required to achieve word learning by applying an auxiliary learning objective on stored trajectories of the agent's experience, in a manner proposed by Jaderberg et al. (2016) (Fig. 2C).[3] In agents with this auxiliary prediction process, the final 4 observations of each episode are saved in a replay buffer and processed offline by the visual and language modules. The concatenation of the output of these modules is then used to predict whether the episode reward was positive, negative or zero. A cross-entropy loss on this prediction is optimised jointly with the agent's A3C loss.

This application of an auxiliary prediction loss can be seen as a rudimentary model of hippocampal replay biased towards rewarding events, a mechanism that is thought to play an important role in both human and animal learning (Gluck & Myers, 1993; Pfeiffer, 2017). The auxiliary loss serves to reinforce the correspondence between visual scenes and words by effectively posing the question *does this word match this view?* This internal question-answering process seems to complement the instruction following, leading to faster word learning at early stages.

Word learning biases

It is widely agreed that children exploit certain labelling biases during early word learning, which serve to constrain the possible referents of novel, ambiguous lexical stimuli (Markman, 1990). Regier (2003) discusses various accounts of how such constraints or biases can emerge naturally from environment signals in connectionist models. A particularly well-studied learning constraint is the *shape bias* (Landau, Smith, & Jones, 1988), whereby infants tend to presume that novel words refer to the shape of an unfamiliar object rather than, say, its colour, size or texture. Our simulated environment permits replication of the original experiments by Landau et al. (1988) that uncovered the shape bias in infants.

During training, the agent learns word meanings in a room containing two objects, one that matches the instruction word (positive reward) and a confounding object that does not (negative reward). Using this method, the agent is taught the meaning of a set C of colour terms, S of shape terms and A of ambiguous terms (in the original experiment, the terms $a \in A$ were the nonsense terms 'dax' and 'riff'). The target referent for a shape term $s \in S$ can be of any colour $c \in C$ and, similarly, the target referent when learning the colours in C can be of any shape. In contrast, the ambiguous terms in A always correspond to objects with a specific colour $c_a \notin C$ and shape $s_a \notin S$ (e.g. 'dax' always refers to a black pencil during training, and neither black nor pencils are observed in any other context).

As the agent learns, we periodically measure its bias by means of test episodes for which no learning takes place. In a test episode, the agent receives an instruction $a \in A$ (e.g. 'dax') and must decide between two objects, o_1, whose shape is s_a and whose colour is $\hat{c} \notin C \cup \{c_a\}$ (e.g. a blue pencil), and o_2, whose shape is $\hat{s} \notin S \cup \{s_a\}$ and whose colour is c_a (e.g. a black fork). Note that in the example neither the colour *blue*

[2] For a video of an agent's behaviour, see `https://tinyurl.com/tcjw5qj`.

[3] Data in this and other learning curves show the best 5 + = SE from 16 replicas with hyperparameters sampled from specific ranges; details available on request.

Figure 3: Development of shape/colour bias as the agent learns. The red curve indicates performance on the training task over time. A positive value for the blue curve reflects a shape bias (max = +10) and negative value reflects a colour bias (max = −10).

nor the shape *fork* are observed by the agent during training. As with the original human experiment, the degree, and colour/shape polarity, of bias in the agent can be measured, as the agent is learning, by its propensity to select o_1 in preference to o_2. Moreover, by varying the size of sets S and C, we can examine how different training regimes affect the bias exhibited by the agent.

Fig. 3 illustrates how a shape/colour bias develops in agents exposed to three different training regimes. The bias is represented by the blue line, which is the mean "score" when +10 is awarded for the object matching the instruction in shape, and -10 for the object matching in colour, over 1000 random test episodes (i.e. a line below zero indicates a propensity to choose objects matching in colour). An agent that is taught exclusively colour words ($|S| = 0$, $|C| = 8$) unsurprisingly develops a strong colour bias. More interestingly, an agent that is taught an equal number of shape and colour terms ($|S| = 8$, $|C| = 8$) also develops a colour bias. In order to induce a (human-like) shape bias, it was necessary to train the agent exclusively on a larger set of ($|S| = 20$, $|C| = 0$) shapes before it began to exhibit a notable shape bias.

It is notable that in the balanced condition our agent architecture (convolutional vision network combined with language instruction embedding) naturally promotes a colour bias. This may be simply because, unlike information pertinent to shapes, the agent has direct access to colour in the RGB stream of pixel input, so that if the environment is balanced, specialising perceptual and grounding mechanisms in favour of colours is a more immediate path to higher returns. Note also that our conclusion differs from that of Ritter, Barrett, Santoro, and Botvinick (2017), who observed a shape bias in convolutional networks trained on ImageNet. Our experiments suggest that this effect is more likely driven by the distribution of training data (the ImageNet data contains many more shape-based than colour-based categories) rather than the underlying convolutional architecture. Indeed, in the present model, it may be that this flexible ability to induce relevant biases facilitates the sudden acceleration of word learning described earlier. As the agent's object recognition and labelling mechanisms specialise (towards shapes, colours or both, as determined by the environment), the space of plausible referents for new words narrows, permitting faster word learning as training progresses.

Indeed, the fact that shape terms occur with greater frequency in typical linguistic environments, for American children at least, can be verified by analysis of the child-directed language corpus Wordbank (Frank, Braginsky, Yurovsky, & Marchman, 2017). Our simulations therefore accord with accounts of the human shape bias that emphasise the role of environmental factors in stimulating the development of such a bias (Regier, 2003). In this view, the human shape bias is not an expression of the default state of underlying perceptual and cognitive mechanisms but rather a product of the prevalence and functional importance of shape categories in the experience of typical infants.

Visualising grounding in action and perception

One compelling aspect of early word learning in humans is infants' ability to make sense of apparently unstructured raw perceptual stimuli. This process requires the learner to induce meaningful extensions for words (when there are limitless potential referents in the environment), and to organise these word meanings in semantic memory. The success of this process has been explained by innate cognitive machinery delimiting conceptual domains, or at least for narrowing the space of possible referents (Marcus, 1999). Alternative accounts, which accord more closely with the learning mechanism presented here, emphasise the capacity of associative learning systems to infer word meanings by exploiting diverse signals in the environment, and bootstrapping currently known words to learn new words more easily (Smith & Yu, 2008; Frank, Tenenbaum, & Fernald, 2013).

First, we visualised the space of word embeddings in an agent trained on words from the different classes shown in

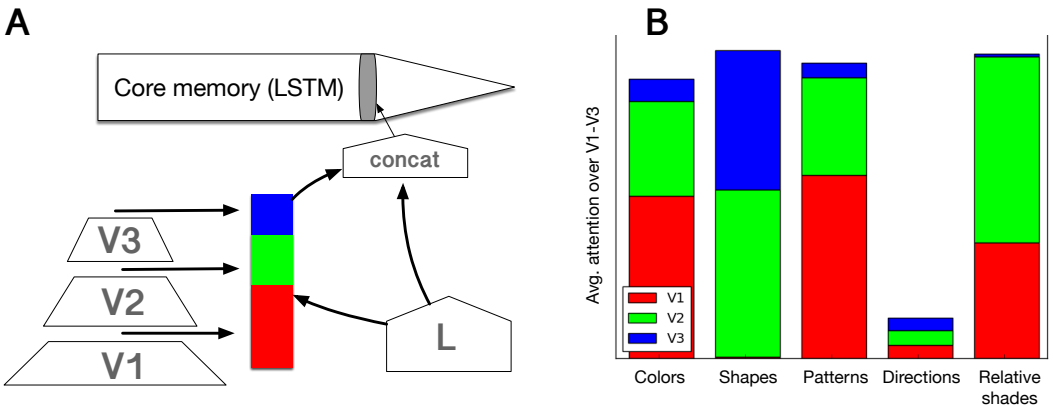

Figure 4: **Online semantic processing in an agent trained with layer-wise visual attention A**: The modified agent network with interactions between language stimuli and all layers of the visual processor. **B**: The attention probability mass allocated to V1-V3 in agents with layer-wise attention trained on words of different types, averaged over all timesteps across 100 episodes after training to convergence on the full set of different words.

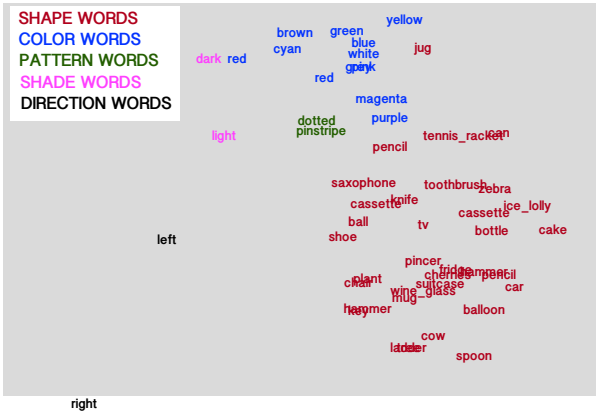

Figure 5: t-SNE visualisation of semantic and syntactic (adjective/noun) classes in the agent's word representation space.

Fig. 5, with experience sampled uniformly over words. We observe that these word classes, which align with both semantic (shape vs. colour) and syntactic (adjective vs. noun) categories, emerge naturally in the embedding space of the agent as it discovered the underlying relationship between words, raw-pixel visual observations of the environment and the 'correct' set of referents as encoded in the environment.

We further explored how this emergent semantic structure manifested itself in processing across the network during an episode. We analysed the trained agent to better understand how it solves the problem of cross-situational word learning in our setting. To do so, we adapted the network to compute weightings for visual field locations at all layers of its visual processing module (a modification we term *layerwise attention*), and measured these weights when agents were trained to understand words of the different types.

More precisely, let e_l be the representation of an instruction word l and $\mathbf{v}_i$ be the output of layer $i = 1,2,3$ of the visual module with dimension $n_i \times n_i \times k_i$, where k_i is the number of feature maps. In the layerwise attention module, the $\mathbf{v}_i$ are first passed through 3 independent linear layers to $\mathbf{v'}_i$ with common final dimension $n_i \times n_i \times K$, such that K is also the dimensionality of e_l. The $\mathbf{v'}_i$ are then stacked into a single tensor T of dimension $d \times K$, where $d = \sum_{i=1}^{3} n_i^2$. T is then multiplied by e_l and passed through a softmax layer to yield a d dimensional discrete probability distribution over all (pixel-like) locations represented in each layer of the visual module $\mathbf{V}$. These values are applied in a weighted sum of the (k_i-dimension) representations returned by each layer before concatenation, as before, with e_l.

By analysing the distribution over spatial locations and visual layers computed by the layerwise attention mechanism, we found that colour and shade words words stimulated activations at the lower levels of the visual-processing module, whereas shape word stimuli activated comparatively more features computed at higher levels (see the red, green and blue bars in Fig 4B, showing activations at levels 1, 2 and 3 of the CNN, respectively). This observation accords with previous analyses of filters in convolutional networks trained for image classification (LeCun, Kavukcuoglu, & Farabet, 2010).

At the mixing layer of the network, we also measured the relative strength of total activation flowing through the visual vs. linguistic pathways for agents trained on different word types, and observed that the direction words were associated with much lower activations from the visual module than other word types. (See the total height of the bars in Fig 4B, which indicates relative activation strength on visual vs. language units, so the higher the bar the more activation on the visual side.) This observation underlines the embodied nature of representation in the agent. Effectively, direction words are grounded in actions to a greater extent than vision, a finding that aligns with cognitive and neuroscientific theories that emphasise the interaction between linguistic semantic representation and sensory-motor processes (Pulvermüller, Mose-

ley, Egorova, Shebani, & Boulenger, 2014).

Simonyan, Vedaldi, and Zisserman (2014) propose a technique for visualising which pixels in an image contribute most to a network's class prediction for that image, using backpropagation to compute the derivative of the model's class score with respect to the image pixels[4]. Here, we can apply a similar technique to a trained agent, but instead computing the derivative of the layerwise-attention probability mass on vision layer V1-V2 with respect to the input image at each timestep. This allows us to render the 'focus' of each of V1-V3 onto the (modified greyscale) visual input as the agent responds to colour or shape instructions, as shown in the video `https://tinyurl.com/tcjw5qj`. A still of this visualization technique is shown in Figure 6.

Discussion and conclusions

Deep reinforcement learning is a comparatively new learning paradigm that is suited to a range of tasks in AI, and has been recently extended to agents conditioned on language input (Oh et al., 2017; Hill et al., 2020). Here we have applied this paradigm to develop an end-to-end, neural-network-based model of cross-situational word learning that can ground word meanings in perception and actions, while relying on few prior assumptions about representation of the visual environment or cognitive states. An appealing aspect of such a model is that the learning process reflects strong interactions between perception, control and language. Such a paradigm may ultimately provide a plausible learning-based computational account for a range of empirical data that emphasize the embodied nature of cognition (Wilson, 2002).

The holistic nature of the simulations also has downsides, however. Since the visual stimuli to our agent is presented as an unstructured array of pixels, it can be challenging to interpret how the agent is making sense of this information. Similarly, since the agent has the freedom to move according to the actions it predicts, as experimenters we lose a degree of control over its visual stimuli across different conditions. Of course, this trade-off between realism and control lies at the heart of experimentation in all human sciences. Our approach affords a degree of novel realism, in that the learning algorithms are instantiated in an agent situated in an environment, but less realistic in its reliance on an abstract simulated world rather than images, videos or care-giver utterances.

Moreover, while our simulations have accounted for some well-known aspects of infant word learning, there are many others that our model in its current guise does not capture. Unlike infants, it is not required to segment the speech stream (Roy & Pentland, 2002), or isolate words from natural (multi-word) child-directed speech (Larsen, Cristia, & Dupoux, 2017). Although word learning in our model gets faster the more it learns, unlike children it is unable to understand a new name for an object after a single experience (so-called fast-mapping) (Xu & Tenenbaum, 2007). The existence of only a single agent in our present environment makes it impossible to consider pragmatic inference or exploit social cues (Frank et al., 2013), and the visual complexity does not match that of the real world (Ritter et al., 2017). Finally, while we have shown that learning can be expedited by offline 'semi-supervised' learning, the predominant learning signal derives from repeated explicit feedback (reward) from following instructions. Such explicit feedback is a frequent experience for language learners in certain cultures, but rare in others (Cristia, Dupoux, Gurven, & Stieglitz, 2017).

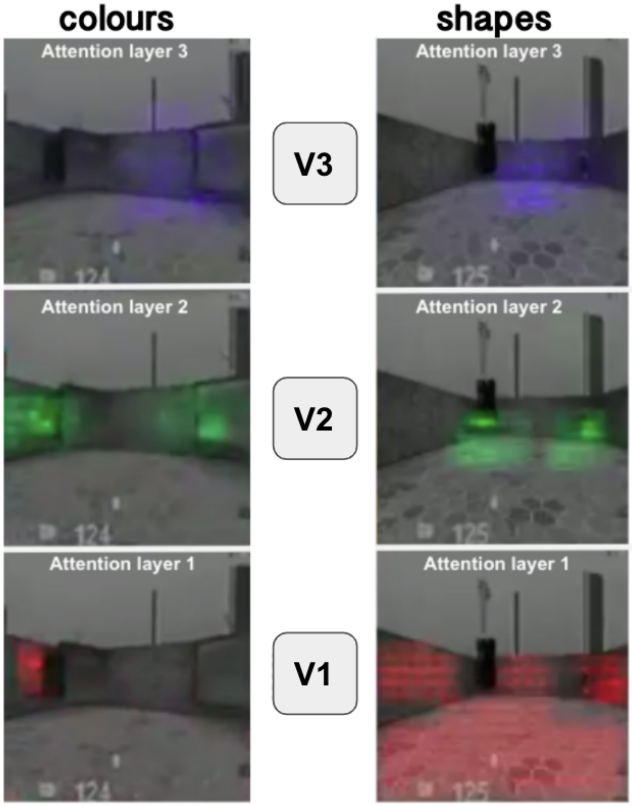

Figure 6: The derivative of the layerwise attention mass allocated to V1, V2 and V3 with respect to input pixels, plotted onto greyscale renderings of the input, at a single timestep of episodes involving a colour or a shape instruction. When performing a colour task, V1 output from the edge of objects (in red) is most important to the model, whereas when performing a shape task, V1 is used to survey the background, while V2 and V3 focus more on the objects in the scene.

[4]In contrast to the standard computation of the derivative of the loss function with respect to the model's weights, computed to determine weight updates during training.

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual Question

Answering. In *Iccv.*

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. *Mechanisms of language acquisition*, 157–193.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Icml.*

Bloom, P. (2000). *How children learn the meanings of words*. The MIT Press.

Cristia, A., Dupoux, E., Gurven, M., & Stieglitz, J. (2017). Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child development*.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

Fernald, A., Thorpe, K., & Marchman, V. A. (2010). Blue car, red car: Developing efficiency in online interpretation of adjectivenoun phrases. *Cognitive Psychology*, *60*(3), 190 - 217.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of child language*, *44*(3), 677–694.

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, *9*(1), 1-24.

Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*, *3*(4), 491–516.

Hill, F., Lampinen, A., Schneider, R., Clark, S., Botvinick, M., McClelland, J. L., & Santoro, A. (2020). Emergent systematic generalization in a situated agent. *ICLR*.

Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D., & Kavukcuoglu, K. (2016). Reinforcement learning with unsupervised auxiliary tasks. In *Iclr.*

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Nips* (pp. 1097–1105).

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, *3*(3), 299–321.

Larsen, E., Cristia, A., & Dupoux, E. (2017). *Relating unsupervised word segmentation to reported vocabulary acquisition.* Open Science Framework.

LeCun, Y., Kavukcuoglu, K., & Farabet, C. (2010). Convolutional networks and applications in vision. In *Icsas* (pp. 253–256).

MacWhinney, B. (1999). *The emergence of language*. Taylor & Francis.

Marcus, G. (1999). *Poverty of the stimulus arguments.* Cambridge, Mass.: MIT Press.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*(1), 57–77.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *Icml* (pp. 1928–1937).

Nazzi, T., & Bertoncini, J. (2003). Before and after the vocabulary spurt: Two modes of word acquisition? *Developmental Science*, *6*(2), 136–142.

Oh, J., Singh, S., Lee, H., & Kohli, P. (2017). Zero-shot task generalization with multi-task deep reinforcement learning. *Proceedings of ICML.*

Pfeiffer, B. E. (2017). The content of hippocampal replay. *Hippocampus*.

Plunkett, K., Sinha, C., Møller, M. F., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? vocabulary growth in children and a connectionist net. *Connection Science*, *4*(3-4), 293–312.

Pulvermüller, F., Moseley, R. L., Egorova, N., Shebani, Z., & Boulenger, V. (2014). Motor cognition–motor semantics: action perception theory of cognition and communication. *Neuropsychologia*, *55*, 71–84.

Regier, T. (2003). Emergent constraints on word-learning: A computational perspective. *Trends in Cognitive Sciences*, *7*(6), 263–268.

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. *Proceedings of ICML.*

Roy, D. K., & Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive science*, *26*(1), 113–146.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Iclr.*

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Tieleman, T., & Hinton, G. (2012). Lecture 6.5-rmsprop. *COURSERA: Neural networks for machine learning*, *4*(2).

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, *9*(4), 625–636.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.